

Summary

Summary

Methodological News contains articles on a range of topics across different areas within the Methodology Division in the ABS. This issue contains two articles, Machine Learning for Editing and Imputation of Big Data, and A New Machine Learning Approach for Entity Resolution.

Machine Learning for Editing and Imputation of Big Data

MACHINE LEARNING FOR EDITING AND IMPUTATION OF BIG DATA

Within statistical agencies, a large amount of human effort and time is put into checking out and correcting data before publication. As we get access to new "big data" sources the amount of data available may increase dramatically, making it infeasible to apply consistent human scrutiny of the data. This could lead to missed opportunities to use existing datasets to inform public decision making.

A new project is looking at how humans and computers can work together to treat potential errors in big data. The computer will identify the types of variables and their relationships in a dataset, using approaches that are robust to common types of errors. It will then suggest rules that identify apparently anomalous values. Human input will be in identifying the right rules to use, rather than in fixing individual cases. The computer will apply these decisions across the dataset and report on the impact of changes.

The project will start by encoding the decisions that humans make easily as we look over the dataset. Unlike a human, the computer can apply these decisions consistently across the whole dataset. The work will identify a set of data features that are useful in understanding items and their relationships, and show how these can be used as a basis for decisions. This basic scrutiny will, at a later stage of the project, be improved and supplemented by machine learning of the important relationships in the data. This allows for imputation of values that were identified as anomalous or missing.

The United Nations Economic Commission for Europe (UNECE) High Level Group for the Modernisation of Official Statistics (HLG-MOS) has recently instituted a Machine Learning project, and the ABS has proposed this project as a pilot study. For this purpose it would be useful to identify a publicly-available dataset to be used in the project, ideally one that is big (with many records and many variables) and that may contain incorrect data. Suggestions are welcome; such a dataset could become part of the ABS pilot study.

For more information, please contact Phil Bell Methodology@abs.gov.au

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

A New Machine Learning Approach for Entity Resolution

A NEW MACHINE LEARNING APPROACH FOR ENTITY RESOLUTION

The problem of entity resolution ("is this thing the same as that thing?") has many applications in ABS work, particularly in the data integration and validation processes.

An internship organised by Australian Postgraduate Research (APR) between December 2018 and April 2019 allowed an intern from the Australian National University (ANU) to work on these issues with the ABS. The intern explored methods for linking business records across time in a simulated population using machine learning.

This kind of linkage is often investigated via *latent variable* methods: we assume that each person, business, or other entity has invisible (latent) features which influence the presence or absence of relationships, and we use the

observable data to simultaneously estimate these features and the rules that determine how such features affect relationships between entities: for instance, if a company employs a high percentage of staff who have previously worked in mining, we might deduce that this company is also in mining.

The project considered several latent variable tensor learning methods including RESCAL, holographic embedding, and complex embedding, before settling on Trans-E. This is a relatively simple method which still performed well at linking business records in the simulated population, and scales well to large data sets. The linkage performance was improved with a “blocking” approach that broke the full-sized problem into smaller, more manageable problems.

We are now working to develop more sophisticated test data sets which will include realistic longitudinal phenomena and relationships between entities (e.g. family members working in the same business) along with complications such as multiple job-holders and enterprise group structures. These will be generated by a simulation approach with no use of confidential inputs, allowing them to be shared freely without any requirement for clearance processes. We expect that this will improve our ability to test and validate machine learning on realistic data, and make it easier to draw on external expertise without needing to share private data.

For more information, please contact Junmei Jing Methodology@abs.gov.au

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

How to Contact Us and Email Subscriber List

HOW TO CONTACT US AND EMAIL SUBSCRIBER LIST

Methodological News features articles and developments in relation to methodology work done within the ABS Methodology Division. By its nature, the work of the division brings it into contact with virtually every other area of the ABS. Because of this, the newsletter is a way of letting all areas of the ABS know of some of the issues we are working on and help information flow. We hope the Methodological Newsletter is useful and we welcome comments.

If you would like to be added to or removed from our electronic mailing list, please contact:

Joey Srinkapaibulaya
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
Belconnen ACT 2617

Email: methodology@abs.gov.au

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

About this Release

Methodological News is a quarterly information bulletin from the Methodology Division. The newsletter features articles and developments in relation to work done with the division.